Beyond the Hox: how widespread is homeobox gene clustering?

PETER W. H. HOLLAND

School of Animal & Microbial Sciences, The University of Reading, UK

(*Accepted 6 April 2001*)

ABSTRACT

The arrangement of Hox genes into physical clusters is fundamental to the patterning of animal body plans, through the phenomenon of colinearity. Other homeobox genes are often described as dispersed, implying they are not arranged into clusters. Contrary to this view, however, two clusters of non-Hox homeobox genes have been reported: the amphioxus ParaHox gene cluster and the *Drosophila* 93D/E cluster (referred to here as the NKL cluster). Here I examine the antiquity of these gene clusters, their conservation and their pattern of evolution in vertebrate genomes. I argue that the ParaHox gene cluster arose early in animal evolution, and duplicated in vertebrates to give the four clusters in human and mouse genomes. The NKL cluster is also ancient, and also duplicated to yield four descendent clusters in mammalian genomes. The NKL and Hox gene clusters were originally chromosomal neighbours, within an ancient and extensive array of at least 30 related homeobox genes. There is no necessary relationship between clustering and colinearity, although it is argued that the ParaHox gene cluster does show modified spatial colinearity. A novel hypothesis for the evolution of ParaHox gene expression in deuterostomes is presented.

Key words: Gene clusters; colinearity; evolution; amphioxus; ParaHox.

INTRODUCTION

Homeobox genes constitute a large and diverse multigene family characterised by a recognizable 180 bp motif encoding the 60 amino acid homeodomain. Many distinct 'classes' of homeobox can be distinguished, of which the Hox genes are the best known. These various homeobox gene classes can be grouped into a few large 'superclasses', based on molecular phylogenetic analyses. For example, in analyses performed by Galliot et al. (1999) and subsequently by Gauchat et al. (2000), the majority of homeobox genes fall into two large and distinct monophyletic assemblages. The first includes Hox genes, together with Cdx, Gsx, Xlox, NK, En, Gbx, Msx, Dlx, Emx and a few other homeobox classes. The above authors refer to this as the 'ANTP class' of genes. I refer to it as the 'ANTP superclass', reserving the term 'class' for the constituent sets of genes, such as Hox, En or Emx. The second large monophyletic grouping identified in these analyses was referred to as the 'PRD class' by Galliot et al.; I refer to this as the 'PRD superclass'. This contains homeobox classes such as Pax, Gsc and Otx. The ANTP superclass and the PRD superclass fall as sister clades in these phylogenetic trees; several more divergent homeobox genes fall outside these two clades, such as the TALE, POU and LIM homeoboxes. In this paper, I will confine my attention to the ANTP superclass. Current evidence suggests this superclass of genes is confined to the animal kingdom.

The best characterised of the ANTP superclass homeobox genes are the Hox genes. These play pivotal roles in anteroposterior patterning of animal embryos, and subsequently adult anatomy, through their roles in coding for and specifying positional information along the main body axis of animals (for reviews see McGinnis & Krumlauf, 1992; Slack et al. 1993). These genes respect spatial cues within embryos, such that distinct positions along the main body axis express different combinations of Hox genes. The genes are arranged in linked clusters or

Correspondence to Professor Peter W. H. Holland, School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AJ, UK. Tel.: + 44 (0) 118 931 8466; fax: + 44 (0) 118 931 6644; e-mail: p.w.h.holland@reading.ac.uk

arrays; in other words, they sit immediately adjacent to each other along a chromosome. A single cluster of Hox genes is present in the genome of all invertebrates studied in detail; in contrast, this single cluster has duplicated to yield multiple Hox gene clusters on separate chromosomes in vertebrates (Schughart et al. 1989; Garcia-Fernàndez & Holland, 1994). The clustering of Hox genes is intimately associated with their regulation, and hence expression, in a way that relates indirectly to anatomy. This unusual association between genome organisation and adult anatomy is mediated through 'colinearity'. The term actually describes two interlinked phenomena: spatial colinearity and temporal colinearity. The former refers to the fact that the spatial order of Hox gene expression along the main anteroposterior axis of developing embryo is the same as the physical order of the genes along the chromosome. In other words, genes at one end of the cluster are expressed (and functional) anteriorly, and each subsequent gene is expressed slightly more posteriorly. Temporal colinearity refers to the observation that the (anterior) genes at one end of the cluster are expressed earlier in development than each subsequent Hox gene along the cluster.

The genomic clustering of Hox genes is striking, but how unique is it? In particular, since Hox genes are closely related to several other homeobox genes in the ANTP superclass, do any of these genes show clustering? If they do, can we deduce the antiquity of homeobox gene clustering? Clustering by itself does not provide a link between genome organisation and anatomy; this is effected through colinearity. Hence, we must also ask if other clustered homeobox genes share the phenomenon of colinearity; if so, in relation to which aspects of embryonic or adult anatomy?

DISCOVERY OF THE PARAHOX GENE CLUSTER

Homeobox genes from classes other than the Hox class were originally described as 'dispersed' homeobox genes, implying that they were scattered around the genome rather than being arranged into clusters. This viewpoint was first challenged by analyses of the NK genes of *Drosophila* (see below) and the ParaHox genes of amphioxus. Genomic clustering of the latter genes first came to light in 1994, during a genomic walk to map the amphioxus Hox gene cluster, carried out by Jordi Garcia-Fernàndez & Holland (1994). In addition to several hundred phage clones containing Hox genes, we isolated one clone that contained two linked homeobox genes, neither of which was a Hox gene. We identified the two amphioxus genes as the amphioxus homologue of *Drosophila caudal* or vertebrate Cdx genes, and the amphioxus homologue of leech *Lox10* or vertebrate *pdx-1* (also called *XlHbox-8*, *ipf-1* or *IDX-1*). These linked genes, therefore, are members of the Cdx and Xlox classes respectively, not the Hox class. It could be argued that two genes do not make a cluster. Together with Nina Brooke, therefore, we extended a genomic walk around these genes, searching for other members of the ANTP superclass. This lead to the discovery that amphioxus possesses a small cluster of three non-Hox homeobox genes: Cdx, Xlox and Gsx (Brooke et al. 1998). The discovery that these genes are physically linked provided conclusive proof that the Hox genes are not the only clustered homeobox genes in chordate genomes.

It is extremely important to determine the age of a gene cluster if one wants to make predictions about its functionality. For example, if a gene cluster has arisen very recently in evolution, by recent tandem gene duplication events, then it may not have a fundamentally important role (mutation and natural selection may simply have not had sufficient time to purge it from the genome) or it may have a function specific to a particular group of closely related species. Three lines of argument, taken together, suggested that the ParaHox gene cluster is extremely ancient, having arisen early in animal evolution. First, at least two of the constituent genes were already known to be phylogenetically widespread and hence ancient. These were Cdx (known, for example, from *Drosophila*, a nematode and several vertebrates) and Xlox (known from leeches and vertebrates). Second, the three homeobox genes are all members of the ANTP superclass and, therefore, are evolutionarily related. It is difficult to envisage how physical clustering of related genes can be derived in evolution from dispersed genes, since this would necessitate a remarkably high frequency and precision of gene shuffling. This implies that the three genes have been linked since their origin and, because at least two of the genes are ancient, clustering must also be ancient. The third line of evidence came from phylogenetic analysis of the protein sequences encoded by ParaHox and Hox genes. This suggested that the ParaHox and Hox gene clusters are paralogues (sisters) that originated by duplication from an ancestral precursor homeobox gene cluster (Brooke et al. 1998). The implication is that any organism with a recognisable Hox gene cluster should also have (or had) a ParaHox gene cluster. Because the Hox gene cluster is conserved across all the 'higher' or bilaterian animals (and probably also the more basal cnidarians; see

Ferrier & Holland, 2001), then it—and by implication the ParaHox gene cluster—must have originated soon after the emergence of animals. Such an ancient origin and maintenance over hundreds of millions of years implies that the ParaHox gene cluster has been selectively retained for functional reasons.

CONSERVATION AND DUPLICATION OF PARAHOX GENES

At the time of writing, physical linkage of ParaHox genes has only been published for amphioxus. This contrasts sharply with the situation for Hox genes, and is rather surprising in the light of the deduction made above that the ancestral condition for these genes is linkage into a gene cluster. It is worth, therefore, revisiting the three lines of evidence that were used to draw this deduction, to see if they have been strengthened or weakened in the 2 y since the ParaHox gene hypothesis was published (Brooke et al. 1998). First, antiquity of the constituent genes. A widespread distribution was noted above for Cdx and Xlox, but such a distribution was not evident for Gsx at the time when the ParaHox hypothesis was formulated. Indeed, the amphioxus Gsx gene described by Brooke et al. (1998) was the first Gsx gene reported outside the vertebrates. This anomaly was remedied by the discovery of a Gsx class gene (*ind*) in *Drosophila* (Weiss et al. 1998); this result confirmed that all three ParaHox genes date to the base of the bilaterian animals. Indeed, recent analyses of cnidarian genes suggest that Cdx and Gsx may be even more widespread, with putative homologues present in cnidarians (for review see Ferrier & Holland, 2001). These data further strengthen the hypothesis that ParaHox genes are ancient. The second line of reasoning was based on the probability that clustering is unlikely to be derived from dispersed genes; this reasoning is still valid. The third line of evidence centred on phylogenetic analyses suggesting that Hox and ParaHox clusters are paralogues (sisters). Several authors have recently re-analysed the relationships between Hox and ParaHox genes, using different molecular phylogenetic methods to those originally used by Brooke et al. (1998). These studies all confirm the original conclusion and imply that any animal with a Hox cluster should have (or had) a ParaHox gene cluster (Finnerty & Martindale, 1999; Ferrier & Holland, 2001).

Although the ParaHox gene cluster is as ancient as the Hox gene cluster, it seems to have been subject to more gene loss; that is, it may be less conserved in animal evolution. For example, the near complete *Drosophila* genome confirms the presence of Cdx and Gsx, but there is no Xlox gene; similarly, the *C. elegans* (nematode) genome has only a Cdx gene (known as *pal-1*) and no Xlox or Gsx (Ruvkun & Oliver, 1998). It is possible, however, that the genomes of these two species are rather atypical for protostome invertebrates. For example, both species have split or rearranged Hox gene clusters, unlike other animals examined. In accord with this conclusion, we have recently been able to clone all three ParaHox genes from two other protostome invertebrate species (Ferrier & Holland, unpublished).

The number and organisation of ParaHox genes in vertebrates is particularly interesting. The most complete information presently available is for the human and mouse genomes (Pollard & Holland, 2000). Three Cdx genes, one Xlox gene and two Gsx genes have been identified in these mammals. In the mouse, the single Xlox gene maps to the same chromosomal position as one of the Cdx genes (*Cdx2*) and one of the Gsx genes (*Gsh2*), on chromosome 5 at 82 cM. These three genes have also been mapped in humans, where they colocalise to 13q12. Analysis of the recently released draft human genome sequence (The Human Genome International Sequencing Consortium, 2001) confirms physical linkage of a ParaHox gene cluster on chromosome 13, although the precise intergenic distances are uncertain due to the unfinished nature of the sequence (www.ensembl.com; release 0.8.0; Pollard & Holland, unpublished analyses). These data from human and mouse, together with the amphioxus results, indicate that a ParaHox gene cluster has been conserved through chordate evolution.

The 3 linked genes do not account for all ParaHox genes in mammals; one additional Gsx and 2 Cdx genes map to other locations. Two contrasting scenarios can be proposed to account for this distribution of genes (Fig. 1). The first is that additional genes were produced by single gene duplication from the ancestral ParaHox gene cluster, resulting in *Cdx1*, *Cdx3/4* and *Gsh1* being dispersed around the genome (left hand side of Fig. 1). In principle, this could occur by tandem gene duplication followed by chromosome transposition or by LINE-mediated retrotransposition, a process shown to be capable of copying genomic DNA to distant locations (Moran et al. 1999). The second scenario is that an ancestral ParaHox gene cluster duplicated in its entirety, to yield four descendant gene clusters. Three of these daughter clusters would then have suffered gene loss resulting in the pattern observed in the human and mouse genomes (right hand side of Fig. 1).

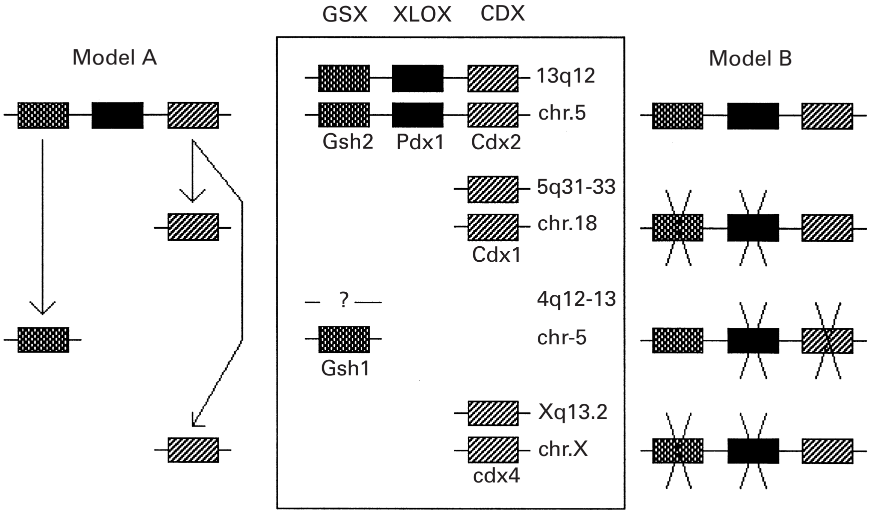

Fig. 1. Chromosomal map positions of human and mouse ParaHox homeobox genes (central panel) and their putative duplication history. The left panel depicts a model involving individual gene duplications; the right panel proposes gene cluster duplication followed by gene loss. The latter is deduced to be correct.

To distinguish between these possibilities, we looked for evidence of paralogy between the human chromosome band positions to which the various ParaHox genes map (Pollard & Holland, 2000). We identified 6 small multigene families that had members mapping to at least 2 of the relevant chromosome band positions. For example, *ED2* (*hidrotic ectodermal dysplasia 2*) maps to 13q12 (close to the human ParaHox gene cluster) while *ED1* (*anhidrotic ectodermal dysplasia 1*) maps close to the singleton *CDX4* at Xq12-13. Similarly, *PDGFRA* (*platelet-derived growth factor receptor*, *alpha polypeptide*) maps close to the inferred position of *GSH1* (inferred by synteny with mouse) at 4q12, while *PDGFRB* (*platelet-derived growth factor receptor*, *beta polypeptide*) maps close to *CDX1* at 5q31-32. These data indicate that the 4 chromosomal regions are evolutionarily related; hence, the second model of gene duplication is correct (right hand side of Fig. 1). We conclude that an ancestral ParaHox gene cluster duplicated in vertebrate evolution to yield 4 descendants in mammals (*A–D*). Three clusters have been degraded by gene loss, but none have been split. This pattern of gene cluster evolution is reminiscent of that described for Hox genes in vertebrates: duplication of the entire gene cluster followed by gene loss.

THE NKL OR 93D/E GENE CLUSTER

In 1989, Kim and Niremberg reported 4 novel homeobox genes from *Drosophila* that are closely related to each other but distinctive from the Hox class (Kim & Niremberg, 1989). These genes were simply named *NK1* to *NK4*, because functions had not been assigned to these genes at the time. Subsequent mutational and complementation studies showed that *NK1* is equivalent to *slo* (*slouch*), *NK2* is *vnd* (*ventral nervous system defective*), *NK3* is *bap* (*bagpipe*) and *NK4* is *tin* (*tinman*); for a review see Harvey (1996). Of particular interest from the present perspective is Kim and Niremberg's finding that 3 of these genes map to the same cytological location in *Drosophila*: 93D/E. The *NK2* gene is the exception, mapping to 1C1-5. Location of related genes at a common cytological position is suggestive of physical clustering produced by tandem gene duplication in evolution, since (as noted above) it is difficult to envisage how similar genes can be assembled from disparate locations. The hypothesis that *NK1*, *NK3* and *NK4* form a gene cluster was subsequently proved by cloning of this genomic region; indeed this revealed that the gene cluster actually contains six tightly linked homeobox genes (Jagla et al. 1994, 1997, 2001; Dear & Rabbitts, 1994). The 3 additional genes are *93Bal* and 2 genes at the *ladybird* locus known as *ladybird early* (*lbe*) and *ladybird late* (*lbl*). Phylogenetic analysis reveals that all six genes are relatively closely related to each other, within the ANTP superclass as previously defined (Gauchat et al. 2000; Pollard & Holland, 2000). This homeobox gene cluster was named the 93D/E cluster, on account of its cytological position in *Drosophila*.

It was stressed above that age of a gene cluster is an important consideration when assessing its functionality and conservation. In the case of the 93D/E gene cluster, all the constituent genes have orthologues in vertebrates, demonstrating that the genes are ex-

Table 1. *Vertebrate orthologues of* Drosophila *93D/E cluster, plus NK2, genes*

Drosophila gene	Vertebrate homologue(s)
NK1 (= slou)	Sax1 (= Nkx1.2), Sax2 (= Nkx1.1)
NK2 (= vnd)	NKX2.1 (= TTF1), NKX2.2, NKX2.4, NKX2.8, Nkx2.9
NK3 (= bap)	NKX3.1 (= NKX3A), Nkx3.2 (= Bapx1)
NK4 (= tin)	NKX2.3, NKX2.5 (= CSX), Nkx2.6, nkx2.7
lbe, lbl	LBX1, LBX2
93Bal	TLX1 (= HOX11), TLX2 (= HOX11L1), Tlx3 (= Hox11L2)

Vertebrate orthologues of genes in the *Drosophila melanogaster* 93D/E cluster as revealed by molecular phylogenetic analyses. *Drosophila NK2* is not part of the 93D/E cluster, but is included because it is an original NK gene cloned by Kim & Niremberg (1989). No vertebrate genes group with fly *NK4* in phylogenetic trees; no fly genes group with vertebrate *NKX2.3*, *NKX2.5*, *Nkx2.6* and *nkx2.7*. We suggest the latter genes are collectively 'cryptic' orthologues of *NK4*, and that rapid sequence divergence has obscured phylogenetic signal. From Pollard & Holland (2000).

tremely ancient (Table 1). There are two slight caveats to this conclusion. First, the *lbe* and *lbl* genes do not have distinct homologues in vertebrates; rather, they seem to be resultant from a tandem gene duplication in the *Drosophila* lineage, with the vertebrate genes *Lbx1* and *Lbx2* descendent from the ancestor to this duplication. Second, molecular phylogenetic analyses do not identify clear homologues of *Drosophila NK4* in vertebrates; however, several lines of evidence suggest that the 'cardiac' group of vertebrate NK genes (*Nkx2.3*, *Nkx2.5*, *Nkx2.6*, *nkx2.7*) are probably 'cryptic' homologues, and that the historical relationship has been obscured by unequal rates of evolution (Harvey, 1996; Pollard & Holland, 2000). Using the same logic as applied in the discussion of ParaHox genes above, the demonstrable antiquity of the 93D/E constituent genes implies that the gene cluster itself is also ancient. For this reason, we prefer to use a more generic term for the cluster that avoids reference to the cytological position in a single species: we refer this cluster as the NK or NKL (NK-like) cluster. For the constituent genes, we use the original *Drosophila* terms for *NK1*, *NK3*, and *NK4*, but prefer the vertebrate terminology for the others (*Tlx* and *Lbx*). In summary, an ancient NKL homeobox gene cluster contained single copies of the *NK1*, *NK3*, *NK4*, *Tlx* and *Lbx* genes (not necessarily in this order).

To assess conservation of the NKL homeobox gene cluster during animal evolution, we collated the chromosomal map positions for the human and mouse NK1, NK3, NK4, Tlx and Lbx class genes, using the NCBI, Jackson laboratory and GDB websites plus BLAST searches of the emerging human genome sequence (Pollard & Holland, 2000). In human, we found map positions for seven genes, but these were located in just 5 chromosomal bands (2p13-14, 4p16, 5q34, 8p21, 10q24). Colocalisation—suggestive of clustering—was detected for *TLX1* and *LBX1*, and also for *TLX2* and *LBX2* (indeed, the homeoboxes of the latter two are just 17.4 kb apart). Considering also that several homologues will have not yet been mapped, and thus the data are incomplete, our interpretation is that these associations are remnants of duplicated NKL gene clusters in the human genome. At first site, a total of 5 chromosomal positions might suggest the presence of five NKL gene clusters in humans. However, analyses of linked (non-homeobox) genes at each of these chromosomal bands suggests that 8p21 and 2p13-14 were actually derived from a single NKL cluster that had been split in vertebrate evolution (Pollard & Holland, 2000).

While collating these data we noted that several other ANTP super class homeobox mapped to these same—or the adjacent—chromosomal bands in human and mouse (Pollard & Holland, 2000). These are the Msx, NK6 (Gtx), Hmx, Emx and Vax class homeobox genes (Fig. 2). For most of these gene classes, no members of the class map to any other chromosomal band (as far as current data indicate), giving a further indication that these colocalisations are significant. For example, there are only two Emx class genes in the human genome: *EMX1* which maps to 2p13-14 (the same band as TLX2 and LBX2), and *EMX2* which maps to 10q25-26, relatively close to *TLX1* and *LBX1* at 10q24. The simplest explanation for these findings is that when the ancestral NKL gene cluster duplicated in vertebrate evolution to yield the four descendent NKL gene clusters, this cluster was adjacent to Msx, NK6, Hmx, Emx and Vax homeobox genes. Thus, a large array of homeobox genes was simultaneously duplicated in vertebrate evolution. Whether these other genes were very tightly linked to the classical NK genes (part of the same cluster) at the time of duplication is not known. Molecular phylogenetic analyses, however, strongly suggest that all these genes were indeed part of a large NKL gene cluster at some time in evolution. For example, Balavoine performed a parsimony analysis of ANTP

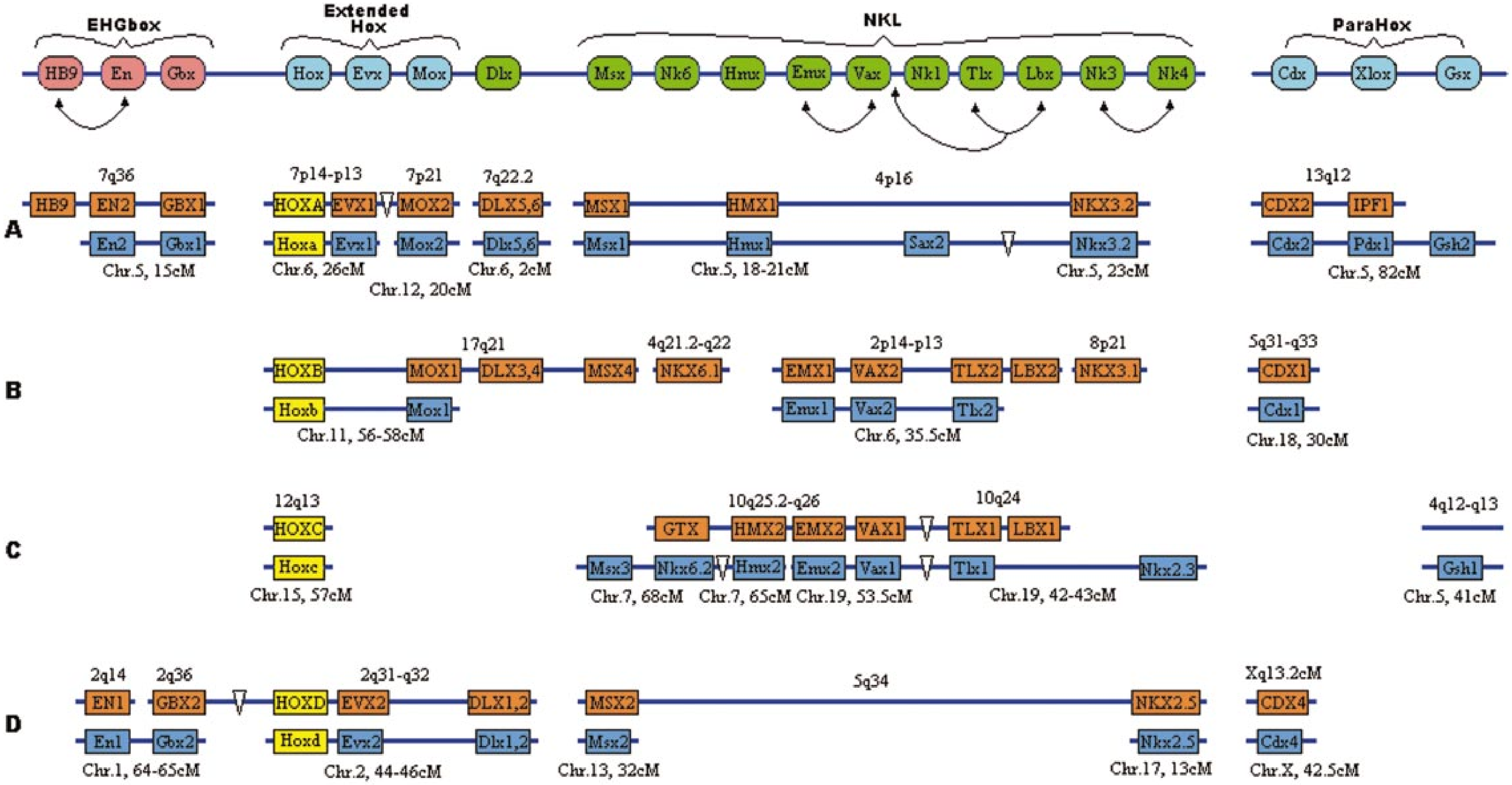

Fig. 2. Chromosomal map positions of human and mouse ANTP superclass homeobox genes below their deduced ancestral linkage arrangements. Each box represents one mapped gene; the upper box being human, the lower mouse. For simplicity the Hox gene clusters are also represented as single boxes. The homeobox linkage groups identified from mapping data are assembled here into eight proposed ancestral chromosomal regions: four containing ParaHox genes and four containing Hox, NKL and others. Double vertical lines indicate breakages of linkage groups in evolution; inverted triangles denote large insertions. The top line indicates the deduced ancestral arrangement of ANTP superclass homeobox genes, before genome duplication in vertebrate evolution. Reproduced from Pollard & Holland (2000).

superclass homeodomains, using carefully calibrated between-site weighting. This revealed an ancient division into a 'Hox-like' grouping and an 'NK-like' grouping (Balavoine, 1996). A similar deep division within the ANTP superclass is evident in the Neighbour Joining trees of Galliot et al. (1999). Importantly, the NK-like grouping includes not just the NK genes, but also Tlx, Lbx and *all* the gene classes that colocalise with these genes in human and mouse genomes (plus just a few others). Consequently, all these gene are close evolutionary relatives, as expected for members of a true gene cluster.

In summary, the gene mapping and molecular phylogenetic analyses suggest that an extensive array of NK-like genes existed in the genomes of a basal bilaterian animal, and has been retained in mammalian genomes. This array included NK1, NK3, NK4, Lbx, Tlx, Emx, Vax, Hmx, NK6 and Msx, and presumably arose by tandem gene duplication. In the lineage leading to *Drosophila*, most of these genes have been secondarily dispersed around the genome, apart from 5 genes (*NK1*, *NK3*, *NK4*, *Tlx*, *Lbx*) that retained very tight linkage. Of these, the *Lbx* gene underwent a subsequent extra tandem gene duplication. On the lineage leading to mammals, the entire and extensive NKL gene array duplicated to give 4 descendent arrays. These have been subsequently affected by gene loss, some chromosomal breakages and quite probably some lateral dispersal along chromosomes, but the 'fingerprints' of their ancestry is still readily detected in mammalian genomes. We reconstruct the minimal constitution of each human NKL array as follows:

NKL-A = MSX1, HMX1, SAX2, NKX3.2
NKL-B (split) = MSX4, EMX1, VAX2, TLX2,
LBX2, NKX3.1
NKL-C (split) = MSX3, NKX6.1, HMX2, EMX2,
VAX1, TLX1, LBX, NKX2.3
NKL-D = MSX2, NKX2.5

A HOMEOBOX GENE MEGA-CLUSTER IN ANIMAL EVOLUTION

The Msx class homeobox genes have been well studied by vertebrate developmental biologists interested in craniofacial development, neural crest cell fate, tooth development and other processes. Two Msx genes, *Msx1* and *Msx2*, have similar expression patterns and putative roles in these processes and may interact in some aspects of tissue morphogenesis (Davidson & Hill, 1991). Ten years ago, an evolutionary study of the Msx class homeobox genes found a third member of this gene class in mammals, designated *Msx3* (Holland, 1991). Expression of this gene during mouse embryogenesis was later characterised in detail (Shimeld et al. 1996; Wang et al. 1996). During BLAST searches of the emerging human genome sequence (Pollard & Holland, 2000), we unexpectedly detected a putative fourth member of this gene class: *MSX4*. Although nothing is known as yet about the expression or function of this gene, its discovery provided a crucial clue to understanding the evolution of homeobox gene clusters.

As outlined above, the *Msx1* and *Msx2* genes map in the putative NKL-A and NKL-D genes clusters of human and mouse, while the *Msx3* gene maps to the NKL-C array of mouse. This leaves NKL-B as the only one of the four arrays not possessing a member of the Msx gene class. We hypothesised, therefore, that the newly discovered human *MSX4* gene was the Msx gene class member assignable to NKL-B (Pollard & Holland, 2000). This gene, however, does not map to the locations of the other NKL-B genes (2p13-14, 4p21-22 and 8p21). Instead, *MSX4* maps to 17q21. This was a surprising finding, since this is the same chromosomal band as one of the Hox gene clusters (HOXB) and several other homeobox genes known to be linked to Hox genes (Mox and Dlx genes; Fig. 2). Our interpretation of this finding is that prior to cluster duplication, the Hox gene cluster and the extensive NKL gene cluster were chromosomal neighbours. After duplication of this locus, chromosomal breakage separated the Hox and NKL clusters between the Hox and Msx genes in copies A, C and D, but instead split the B array between *MSX4* and the rest of the NKL array. This hypothesis also explains the anomalous observation that the Dlx class homeobox genes are NK-like on the basis of sequence, yet linked to Hox genes rather than NKL genes in terms of chromosomal location. In the scenario proposed by Pollard & Holland (2000), the Dlx genes were ancestrally NKL genes but, like *MSX4*, chromosomal breakages have left them stranded with the Hox genes.

The wider implication of this finding is that numerous ANTP superclass homeobox genes were ancestrally found within a single chromosomal region. Indeed, the above analyses suggest that minimally 27 different ANTP superclass genes were linked into an array before cluster duplication in the vertebrate lineage. These were 13 Hox genes, two Dlx genes, Evx, Mox, Msx, NK6, Hmx, Emx, Vax, NK1, NK3, Lbx, Tlx and NK4. Indeed, for reasons explained in Pollard and Holland (2000), the actual number is slightly higher at 30, since analyses of linked non-homeobox genes indicate that En, Gbx and Mnx (= HB9) classes were also linked. It is not yet possible to deduce the precise ancestral physical order of these 30 homeobox genes, nor the relative distances between them. We can infer, however, that the Hox genes formed a tight cluster within this 'mega-array', as did at least 5 of the NKL genes. The other genes may have been tightly associated with one or other cluster, or may have already dispersed across this genomic region, becoming intermingled with numerous non-homeobox genes. We can also infer that the duplication of the megacluster occurred after the divergence of the vertebrate lineage from the cephalochordates (amphioxus). This is clear because amphioxus (the closest invertebrate relative of the vertebrates) has a single Hox gene cluster, and most likely single En, Mnx, Msx, NK1, NK3 and NK4 genes (Garcia-Fernandez & Holland, 1994; Holland, L. Z. et al. 1997; Ferrier et al. 2001; Sharman et al. 1999; Luke & Holland, unpublished data).

DEVELOPMENTAL ROLES OF 'NOVEL' CLUSTERS

At the start of this article, I posed 3 principal questions concerning the clustering of homeobox genes. First, how unique is Hox gene clustering? Second, what is the antiquity of homeobox gene clustering? Third, do all clustered homeobox genes display colinearity? In answer to the first question, 2 examples of non-Hox homeobox gene clusters were discussed: the ParaHox gene cluster of amphioxus and the 93D/E (NKL) cluster of *Drosophila*. In each case, gene mapping evidence suggests that the gene cluster has been conserved in human and mouse genomes, despite cluster duplication. Furthermore, we have recently produced a partial physical map for the NKL gene cluster in amphioxus, confirming tight linkage of these genes in a chordate (Luke & Holland, unpublished data). Hence, Hox genes are not the only clustered homeobox genes, and it is incorrect to refer to non-Hox class homeobox genes as 'dispersed' homeobox genes. It may be relevant that the Hox, ParaHox and NK genes, plus all the NK-related genes that map close to the NKL and Hox gene clusters (Msx, Emx, Dlx, etc.) are members of the ANTP superclass of homeobox genes. None of the examples discussed here are members of the PRD superclass or more divergent categories of homeobox genes. This raises the possibility that clustering is a phenomenon characteristic of the ANTP superclass. This hypothesis suggests that clustering may be detected for other ANTP superclass genes, besides the Hox, ParaHox and NKL genes already discussed. Such genes include the En, Gbx, Mnx, NK2 and BarH genes. We are currently testing this prediction.

The question of antiquity was also addressed. Because a cluster of related genes cannot be easily assembled secondarily from dispersed genes, it follows that the antiquity of clustering is the same as the antiquity of the individual genes. Orthologues of specific Hox, ParaHox and NKL genes have been identified right across the diversity of bilaterian animal phyla, and—in some cases—within the diploblastic animals such as jellyfish, hydroids, coral and sea anemones (see Gauchat et al. 2000; Ferrier & Holland, 2001). None of these genes have been identified outside the animal kingdom, however, and data from one of the most basal animal lineages, sponges, is hard to interpret. The implication is that an ancestral ANTP superclass homeobox gene arose in the genome of a very basal animal (or possibly ancestor to the animals). This underwent a series of tandem gene duplications, to give precursors of Hox, NKL and several other genes. These precursor genes continued to duplicate in tandem to give an extensive array of ANTP superclass homeobox genes in early animals (at least 30 genes). Sometime during this process, part or all of this array duplicated such that the ParaHox gene cluster was created at a distinct genomic location. All these events preceded the divergence of the lineages leading to arthropods and to chordates. The extensive array of 30 or more ANTP superclass genes was subsequently broken in several animal genomes (including *Drosophila*, nematode and mammalian), although we infer that it was intact in the common ancestor of amphioxus and vertebrates.

The question of colinearity is harder to answer with current data. Specifically, we need to resolve whether colinearity is as ancient as clustering, or whether it is a derived characteristic acquired by the Hox gene cluster. It is premature to assess whether the NKL gene cluster displays colinearty. Jagla et al. (2001) have suggested that a form of temporal colinearity is evident in the expression of *Drosophila* NKL cluster genes in developing mesoderm, but there is little evidence of spatial colinearity. In vertebrates, there is insufficient information about expression of these genes in early embryos to assess this possibility. Furthermore, duplication of the NKL gene cluster and the subsequent extensive gene losses may have complicated detection of colinearity in vertebrates.

Brooke et al. (1998) examined the expression of the three ParaHox genes in amphioxus, and concluded that this gene cluster does display spatial colinearity, although it is less obvious than seen for Hox genes. The 'posterior' ParaHox gene Cdx (related to posterior Hox genes) is expressed in caudal tissues of amphioxus embryos, starting at gastrulation. A

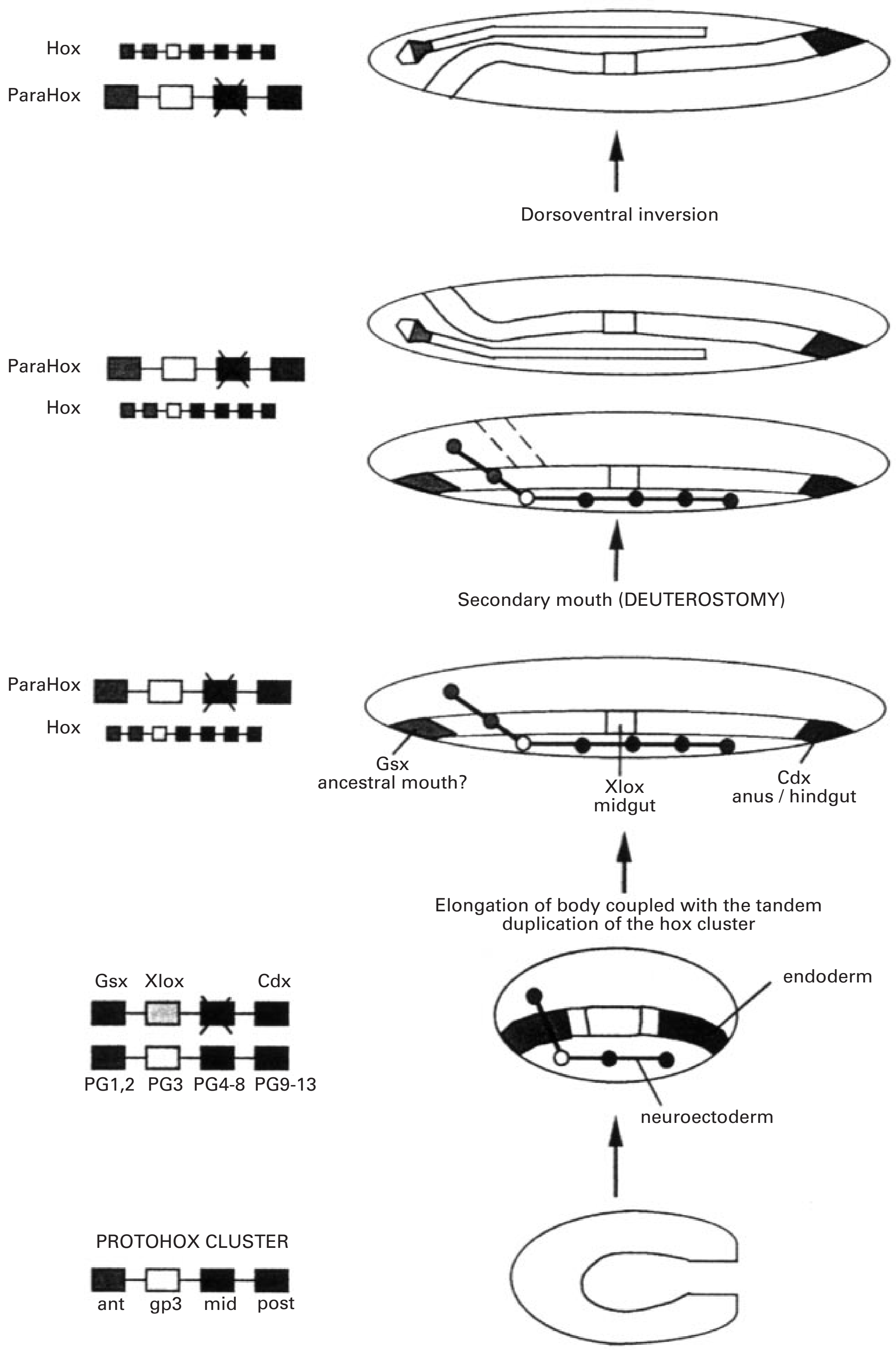

Fig. 3. Hypothetical model for the evolution of Hox and ParaHox gene expression patterns during animal evolution. In the scenario, both the Hox and ParaHox gene clusters displayed complete spatial colinearity in a basal bilaterian animal, after their origin by duplication from an ancestral ProtoHox gene cluster. This colinearity was in different tissues or germ layers; ParaHox in the endoderm, Hox in the neurectoderm. In the deuterostome lineage, a new secondary mouth evolved and dorsoventral axis inversion occurred. The Gsx gene was freed from a patterning role in the 'old' mouth and was redeployed in brain development. Reproduced from Brooke (1999).

dominant site of expression is the posterior endoderm, although there is also persistent expression in posterior neural tissue. Similar expression has been reported for Cdx genes in vertebrates and *Drosophila* (Duprey et al. 1988; Calleja et al. 1996). The central ParaHox gene, Xlox, is also expressed in endoderm, but in this case the domain of expression is rather central in the animal, with clear rostral and caudal boundaries. This expression site overlaps with that described for the amphioxus homologue of insulin and IGF (Holland, P.W.H. et al. 1997), suggesting this homeobox gene may mark the amphioxus homologue of the pancreas. Interestingly, the single mammalian Xlox gene is required for development of the pancreas and anterior duodenum, and acts as a transcription factor in the pancreas (Jonsson et al. 1994; Offield et al. 1996). Even in leeches, a taxon very far removed phylogenetically from the chordates, the Xlox genes are expressed in midgut tissues (Wysocka-Diller et al. 1995).

With the expression of Cdx and Xlox in mind, it might be expected that the third ParaHox gene (related to anterior Hox genes) would be expressed in anterior endoderm. If this were the case, the three genes would display a neat colinearity in the gut. In fact, the amphioxus Gsx gene is expressed in the cerebral vesicle (the homologue of the vertebrate forebrain), mirroring the expression and function reported for Gsx genes in mouse (Li et al. 1996; Szucsik et al. 1997) and expression in fish (Deschet et al. 1998). In our original paper describing the expression of amphioxus Gsx (Brooke et al. 1998), we speculated that this discrepancy might reflect the fact that the major brain structure affected in *Gsh-1* mutant mice, the adenohypophysis, receives developmental input from the oral cavity. Hence, perhaps Gsx functions in anterior gut during early development, but these cells then contribute to brain structures. We now feel this explanation is invalid. One reason for this doubt is based on the elegant work of Kawamura and Kikuyama (1992) demonstrating that the adenohypophysis does not derive from the oral cavity, as long believed, but rather from migrating rostral ectoderm cells that associate only transiently with oral tissue. Hence, the adenohypothysis is unlikely to have an evolutionary link with the mouth.

There is an alternative, and more feasible, hypothesis to account for the anomolous expression pattern of Gsx. First, assume that an association between Gsx and anterior gut development existed in basal animals, when the ParaHox gene cluster originated and became distinct from the Hox gene cluster. At this stage, the three ParaHox genes would have patterned anterior, middle and posterior gut, in a colinear manner. This pattern is retained by Cdx and Xlox of amphioxus and vertebrates, but not by Gsx. The reason may be that these organisms are deuterostomes, and as such their lineage is believed to have undergone a radical alteration of anterior gut formation during evolution. Classical embryological comparisons suggest that the primary mouth was lost in the deuterostome lineage, and a new mouth invented (deuterostome = second mouth). In this scenario, we would not expect Gsx to continue to have a function in the amphioxus or vertebrate oral cavity, since this bears no relation to the ancestral mouth. During this evolutionary modification, Gsx would have been freed form its role in anterior gut formation and may have been recruited to a new role in deuterostomes (e.g. in the brain; Fig. 3). This hypothesis is highly speculative, but at least can be tested by analysis of ParaHox genes in protostome invertebrates, since these have not undergone loss and reinvention of the mouth. The prediction is that protostomes, particularly spiralian protostomes, would retain show colinear expression of ParaHox gene expression in the gut: Gsx patterning the mouth, Xlox the midgut, and Cdx the hindgut and anus. This prediction is being tested.

ACKNOWLEDGEMENTS

I thank Sophie Pollard, Graham Luke, Jordi Garcia-Fernàndez, Nina Brooke and David Ferrier for their contributions to the research discussed in this article, and Nina Brooke for permission to reproduce the deuterostome evolutionary scheme from her PhD thesis. The author's work is funded by the BBSRC.

REFERENCES

BALAVOINE G (1996) *Origine et diversification des complexes de gènes Hox: inventaire de ces gènes chez la planaire Polycelis nigra*. PhD thesis, Université Paris-Sud.

BROOKE NM (1999) *The origin and evolution of Hox-like genes: insights from amphioxus*. PhD thesis, University of Reading.

BROOKE NM, GARCIA-FERNÀNDEZ J, HOLLAND PWH (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392**, 920–922.

CALLEJA M, MORENO E, PELAZ S, MORATA G (1996) Visualization of gene expression in living adult *Drosophila*. *Science* **274**, 252–255.

DAVIDSON DR, HILL RE (1991) Msh-like genes: a family of homeobox genes with wide-ranging expression during vertebrate development. *Seminars in Developmental Biology* **2**, 405–412.

DEAR TN, RABBITTS TH (1994) *A Drosophila melanogaster* homolog of the T-cell oncogene *HOX11* localises to a cluster of homeobox genes. *Gene* **141**, 225–229.

DESCHET K, BOURRAT F, CHOURROUT D, JOLY JS (1998) Expression domains of the medaka (*Oryzias latipes*) Ol-Gsh 1 gene are reminiscent of those of clustered and orphan homeobox genes. *Development*, *Genes and Evolution* **208**, 235–244.

DUPREY P, CHOWDHURY K, DRESSLER GR, BALLING R, SIMON D, GUENET J et al. (1988) A mouse gene homologous to the *Drosophila* gene *caudal* is expressed in epithelial cells from the embryonic intestine. *Genes and Development* **2**, 1647–1654.

FERRIER DEK, HOLLAND PWH (2001) Ancient origins of the Hox gene cluster. *Nature Reviews Genetics* **2**, 33–38.

FERRIER DEK, BROOKE NM, PANOPOULOU G, HOLLAND PWH (2001) The Mnx homeobox gene class defined by HB9, MNR2 and amphioxus AmphiMnx. *Development*, *Genes and Evolution*, **211**, 103–107.

FINNERTY JR, MARTINDALE MQ (1999) Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. *Evolution and Development* **1**, 16–23.

GALLIOT B, DE VARGAS C, MILLER D (1999) Evolution of homeobox genes: Q50 Paired-like genes founded the paired class. *Development*, *Genes and Evolution* **209**, 186–197.

GARCIA-FERNÀNDEZ J, HOLLAND PWH (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370**, 563–566.

GAUCHAT D, MAZET F, BERNEY C, SCHUMMER M, KREGER S, PAWLOWSKI J et al. (2000) Evolution of Antp-class genes and differential expression of *Hydra* Hox/paraHox genes in anterior patterning. *Proceedings of the National Academy of Sciences of the USA* **97**, 4493–4498.

HARVEY RP (1996) NK-2 homeobox genes and heart development. *Developmental Biology* **178**, 203–216.

HOLLAND LZ, KENE M, WILLIAMS NA, HOLLAND ND (1997) Sequence and embryonic expression of the amphioxus engrailed gene (*AmphiEn*): the metameric pattern of transcription resembles that of its segment-polarity homolog in *Drosophila*. *Development* **124**, 1723–1732.

HOLLAND PWH (1991) Cloning and evolutionary analysis of msh-like homeobox genes from mouse, zebrafish and ascidian. *Gene* **98**, 253–257.

HOLLAND PWH, PATTON SJ, BROOKE NM, GARCIA-FERNÀNDEZ J (1997) Genetic patterning of ectoderm and endoderm in amphioxus: from homeobox genes to hormones. In *Advances in Comparative Endocrinology* (ed. Kawashima S, Kikuyama S), pp. 247–252. Bologna: Moduzzi Editore.

JAGLA K, STANCEVA I, DRETZEN G, BELLARD F, BELLARD M (1994) A distinct class of homeodomain proteins is encoded by two sequentially expressed *Drosophila* genes from the 93D/E cluster. *Nucleic Acids Research* **22**, 1202–1207.

JAGLA K, JAGLA T, HEITZLER P, DRETZEN G, BELLARD F, BELLARD M (1997) *ladybird*, a tandem of homeobox genes that maintain wingless expression in terminal and dorsal epidermis of the *Drosophila* embryo. *Development* **124**, 91–100.

JAGLA K, BELLARD M, FRASCH M (2001) A cluster of *Drosophila* homeobox genes involved in mesoderm differentiation programs. *BioEssays* **23**, 125–133.

JONSSON J, CARLSSON L, EDLUND T, EDLUND H (1994) Insulin-promoter-factor-1 is required for pancreas development in mice. *Nature* **371**, 606–609.

KAWAMURA K, KIKUYAMA S (1992) Evidence that hypophysis and hypothalamus constitute a single entity from the primary stage of histogenesis. *Development* **115**, 1–9.

KIM Y, NIRENBERG M (1989) *Drosophila* NK-homeobox genes. *Proceedings of the National Academy of Sciences of the USA* **86**, 7716–7720.

LI H, ZEITLER PS, VALERIUS MT, SMALL K, POTTER SS (1996) *Gsh-1*, an orphan Hox gene, is required for normal pituitary development. *EMBO Journal* **15**, 714–724.

McGINNIS W, KRUMLAUF R (1992) Homeobox genes and axial patterning. *Cell* **68**, 283–302.

MORAN JV, DEBERARDINIS RJ, KAZAZIAN HH (1999) Exon shuffling by L1 retrotranspositon. *Science* **283**, 1530–1534.

OFFIELD MF, JETTON TL, LABOSKY PA, RAY M, STEIN RW, MAGNUSON MA et al. (1996) *Pdx-1* is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122**, 983–995.

POLLARD SL, HOLLAND PWH (2000) Evidence for 14 homeobox gene clusters in human genome ancestry. *Current Biology* **10**, 1059–1062.

RUVKUN G, OLIVER O (1998) The taxonomy of developmental control in *Caenorhabditis elegans*. *Science* **282**, 2033–2041.

SCHUGHART K, KAPPEN C, RUDDLE FH (1989) Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proceedings of the National Academy of Sciences of the USA* **86**, 7067–7071.

SHARMAN AC, SHIMELD SM, HOLLAND PWH (1999) An amphioxus Msx gene expressed predominantly in the dorsal neural neural tube. *Development*, *Genes and Evolution* **209**, 260–263.

SHIMELD SM, McKAY IJ, SHARPE PT (1996) The murine homeobox gene *Msx-3* shows highly restricted expression in the developing neural tube. *Mechanisms of Development* **55**, 201–210.

SLACK JMW, HOLLAND PWH, GRAHAM CF (1993) The zootype and the phylotypic stage. *Nature* **361**, 490–492.

SZUCSIK JC, WITTE DP, LI H, PIXLEY SK, SMALL KM, POTTER SS (1997) Altered forebrain and hindbrain development in mice mutant for the *Gsh-2* homeobox gene. *Developmental Biology* **191**, 230–242.

THE HUMAN GENOME INTERNATIONAL SEQUENCING CONSORTIUM (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

WANG WD, CHEN XW, XU H, LUFKIN T (1996) Msx3: a novel murine homologue of the Drosophila msh homeobox gene restricted to the dorsal embryonic central nervous system. *Mechanisms of Development* **58**, 203–215.

WEISS JB, VON OHLEN T, MELLERICK DM, DRESSLER G, DOE CQ, SCOTT MP (1998) Dorsoventral patterning in the *Drosophila* central nervous system: *the intermediate neuroblasts defective* homeobox gene specifies intermediate column identity. *Genes and Development* **12**, 3591–3602.

WYSOCKA-DILLER J, AISEMBERG GO, MACAGMO ER (1995) A novel homeobox cluster expressed in repeated structures of the midgut. *Developmental Biology* **171**, 439–447.